Taming the data deluge by enriching AI algorithms with new processors

November 1 2021, by Sandi Miller

The vision of A3D3 is to establish a tightly coupled organization of domain scientists, computer scientists, and engineers that unite three core components which are essential to achieve real-time AI to transform science: AI techniques, computing hardware, and scientific applications. Credit: A3D3

An oncoming tsunami of data threatens to overwhelm huge data-rich research projects in such areas that range from the tiny neutrino to an exploding supernova, as well as the mysteries deep within the brain.

When LIGO picks up a gravitational-wave signal from a distant collision of black holes and neutron stars, a clock starts ticking for capturing the

earliest possible light that may accompany them: time is of the essence in this race. Data collected from electrical sensors monitoring brain activity are outpacing computing capacity. Information from the Large Hadron Collider (LHC)'s smashed particle beams will soon exceed 1 petabit per second.

To tackle this approaching data bottleneck in real-time, a team of researchers from nine institutions led by the University of Washington, including MIT, has received $15 million in funding to establish the Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute. From MIT, the research team includes Philip Harris, assistant professor of physics, who will serve as the deputy director of the A3D3 Institute; Song Han, assistant professor of electrical engineering and computer science, who will serve as the A3D3's co-PI; and Erik Katsavounidis, senior research scientist with the MIT Kavli Institute for Astrophysics and Space Research.

Infused with this five-year Harnessing the Data Revolution Big Idea grant, and jointly funded by the Office of Advanced Cyberinfrastructure, A3D3 will focus on three data-rich fields: multi-messenger astrophysics, high-energy particle physics, and brain imaging neuroscience. By enriching AI algorithms with new processors, A3D3 seeks to speed up AI algorithms for solving fundamental problems in collider physics, neutrino physics, astronomy, gravitational-wave physics, computer science, and neuroscience.

"I am very excited about the new Institute's opportunities for research in nuclear and particle physics," says Laboratory for Nuclear Science Director Boleslaw Wyslouch. "Modern particle detectors produce an enormous amount of data, and we are looking for extraordinarily rare signatures. The application of extremely fast processors to sift through these mountains of data will make a huge difference in what we will measure and discover."

The seeds of A3D3 were planted in 2017, when Harris and his colleagues at Fermilab and CERN decided to integrate real-time AI algorithms to process the incredible rates of data at the LHC. Through email correspondence with Han, Harris' team built a compiler, HLS4ML, that could run an AI algorithm in nanoseconds.

"Before the development of HLS4ML, the fastest processing that we knew of was roughly a millisecond per AI inference, maybe a little faster," says Harris. "We realized all the AI algorithms were designed to solve much slower problems, such as image and voice recognition. To get to nanosecond inference timescales, we recognized we could make smaller algorithms and rely on custom implementations with Field Programmable Gate Array (FPGA) processors in an approach that was largely different from what others were doing."

A few months later, Harris presented their research at a physics faculty meeting, where Katsavounidis became intrigued. Over coffee in Building 7, they discussed combining Harris' FPGA with Katsavounidis's use of machine learning for finding gravitational waves. FPGAs and other new processor types, such as graphics processing units (GPUs), accelerate AI algorithms to more quickly analyze huge amounts of data.

"I had worked with the first FPGAs that were out in the market in the early '90s and have witnessed first-hand how they revolutionized front-end electronics and data acquisition in big high-energy physics experiments I was working on back then," recalls Katsavounidis. "The ability to have them crunch gravitational-wave data has been in the back of my mind since joining LIGO over 20 years ago."

Two years ago they received their first grant, and the University of Washington's Shih-Chieh Hsu joined in. The team initiated the Fast Machine Lab, published about 40 papers on the subject, built the group to about 50 researchers, and "launched a whole industry of how to

explore a region of AI that has not been explored in the past," says Harris. "We basically started this without any funding. We've been getting small grants for various projects over the years. A3D3 represents our first large grant to support this effort."

"What makes A3D3 so special and suited to MIT is its exploration of a technical frontier, where AI is implemented not in high-level software, but rather in lower-level firmware, reconfiguring individual gates to address the scientific question at hand," says Rob Simcoe, director of MIT Kavli Institute for Astrophysics and Space Research and the Francis Friedman Professor of Physics. "We are in an era where experiments generate torrents of data. The acceleration gained from tailoring reprogrammable, bespoke computers at the processor level can advance real-time analysis of these data to new levels of speed and sophistication."

The huge data from the Large Hadron Collider

With data rates already exceeding 500 terabits per second, the LHC processes more data than any other scientific instrument on earth. Its future aggregate data rates will soon exceed 1 petabit per second, the biggest data rate in the world.

"Through the use of AI, A3D3 aims to perform advanced analyses, such as anomaly detection, and particle reconstruction on all collisions happening 40 million times per second," says Harris.

The goal is to find within all of this data a way to identify the few collisions out of the 3.2 billion collisions per second that could reveal new forces, explain how dark matter is formed, and complete the picture of how fundamental forces interact with matter. Processing all of this information requires a customized computing system capable of interpreting the collider information within ultra-low latencies.

"The challenge of running this on all of the 100s of terabits per second in real-time is daunting and requires a complete overhaul of how we design and implement AI algorithms," says Harris. "With large increases in the detector resolution leading to data rates that are even larger the challenge of finding the one collision, among many, will become even more daunting."

The brain and the universe

Thanks to advances in techniques such as medical imaging and electrical recordings from implanted electrodes, neuroscience is also gathering larger amounts of data on how the brain's neural networks process responses to stimuli and perform motor information. A3D3 plans to develop and implement high-throughput and low-latency AI algorithms to process, organize, and analyze massive neural datasets in real time, to probe brain function in order to enable new experiments and therapies.

With Multi-Messenger Astrophysics (MMA), A3D3 aims to quickly identify astronomical events by efficiently processing data from gravitational waves, gamma-ray bursts, and neutrinos picked up by telescopes and detectors.

The A3D3 researchers also include a multi-disciplinary group of 15 other researchers, including project lead the University of Washington, along with Caltech, Duke University, Purdue University, UC San Diego, University of Illinois Urbana-Champaign, University of Minnesota, and the University of Wisconsin-Madison. It will include neutrinos research at Icecube and DUNE, and visible astronomy at Zwicky Transient Facility, and will organize deep-learning workshops and boot camps to train students and researchers on how to contribute to the framework and widen the use of fast AI strategies.

"We have reached a point where detector network growth will be

transformative, both in terms of event rates and in terms of astrophysical reach and ultimately, discoveries," says Katsavounidis. "'Fast' and 'efficient' is the only way to fight the 'faint' and 'fuzzy' that is out there in the universe, and the path for getting the most out of our detectors. A3D3 on one hand is going to bring production-scale AI to gravitational-wave physics and multi-messenger astronomy; but on the other hand, we aspire to go beyond our immediate domains and become the go-to place across the country for applications of accelerated AI to data-driven disciplines."

More information: Alec Gunny et al, Hardware-accelerated Inference for Real-Time Gravitational-Wave Astronomy. arXiv:2108.12430v1 [gr-qc], arxiv.org/abs/2108.12430

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Taming the data deluge by enriching AI algorithms with new processors (2021, November 1) retrieved 3 October 2023 from https://techxplore.com/news/2021-11-deluge-enriching-ai-algorithms-processors.html